ARMY RESEARCH LABORATORY

A Comparison of Various Methods Used to Determine the Sample Size Requirements for Meeting a 90/90 Reliability Specification

by David W. Webb

ARL-TR-5468 **March 2011**

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

ARL-TR-5468 **March 2011**

A Comparison of Various Methods Used to Determine the Sample Size Requirements for Meeting a 90/90 Reliability Specification

David W. Webb
Weapons and Materials Research Directorate, ARL

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
March 2011	Final	October 2010 – November 2010

4. TITLE AND SUBTITLE
A Comparison of Various Methods Used to Determine the Sample Size Requirements for Meeting a 90/90 Reliability Specification

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
David W. Webb

5d. PROJECT NUMBER
FPHY10

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-WML-A
Aberdeen Proving Ground, MD 21005-5066

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-5468

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
A common rule of thumb used in binomial-outcome reliability studies is that 22 trials with zero failures are needed to conclude a 90% minimal success rate with 90% confidence. This report explores the origins of this rule and shows how it is overly conservative. Alternative methods for constructing lower confidence bounds for binomial proportions are described which allow the 90/90 reliability criteria to be achieved with as few as 13 trials.

15. SUBJECT TERMS
confidence interval, coverage probability, Clopper-Pearson, Wilson score, Jeffrey

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			David W. Webb
Unclassified	Unclassified	Unclassified	UU	26	19b. TELEPHONE NUMBER (*Include area code*) 410-278-7014

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures iv

List of Tables v

Acknowledgments vi

1. Introduction 1

2. Interval Estimation of Binomial Proportions 1

2.1 The Clopper-Pearson Method 2

2.2 Wilson Score Method 2

2.3 Jeffreys Method 3

3. Comparison of Different LCB Methods 4

4. Sample Size Requirements for Nonzero Failure Testing 7

4.1 Wilson Score Method With One Failure 7

4.2 Wilson Score Method With Two Failures 7

4.3 Jeffreys Method With One Failure 7

4.4 Jeffreys Method With Two Failures 8

5. Multistage Sampling Plans 8

5.1 Operating Characteristic Curves 9

5.2 Sample Size Expectation 10

6. Summary 14

7. References 15

Distribution List 16

List of Figures

Figure 1. A comparison of coverage probabilities vs. true binomial parameter for the nominal 90% LCB. 5

Figure 2. Mean coverage probability as a function of *N:* for the nominal 90% Clopper-Pearson (E), Wilson score (W), and Jeffreys (J) LCBs, when *p* has (a) a uniform distribution on (.7, 1) and (b) a beta distribution with parameters 13.6 and 2.4.....6

Figure 3. Sampling protocol charts for a 90/90 (a) one-stage test, (b) two-stage test, and (c) three-stage test, each using the Wilson score method for estimating the LCB. Travel starts at (0,0) and moves one unit to the right with each trial, and an additional unit upward if that trial is a failure. If the current test "position" is on a dot, then testing continues. If the current test position is on a square, then testing stops without meeting the 90/90 reliability specification. If the current test position is on a circle, then testing stops with the 90/90 reliability specification satisfied. 9

Figure 4. Relationship between percentage of defective items in the population and the probability of meeting the 90/90 reliability specification using (a) Wilson score and (b) Jeffreys methods of computing the LCB for reliability. 11

Figure 5. Percentage of defective items in the population vs. expected number of samples under several sampling protocols using (a) the Wilson score and (b) Jeffreys methods of constructing LCBs. 13

List of Tables

Table 1. Minimum number of trials required to test for a 90/90 reliability standard. 8

Acknowledgments

The author gratefully acknowledges Mr. Zachary Zimmer of the U.S. Army Evaluation Center for recommending the article by Cai which prompted the inclusion of the Jeffreys method in this report.

The author also extends his appreciation for the technical review of Mr. Benjamin Flanders of the U.S. Army Research Laboratory. His insightful comments and suggestions have enhanced the clarity and overall quality of the report.

1. Introduction

For certain reliability studies, the objective is to conduct trials of a pass-fail system under identical conditions with the desired objective of showing that the reliability of the system can be shown to be at some minimum specification level with a predetermined level of confidence. A natural question is "How many trials are necessary to meet this specification?"

In mathematical language, we seek the sample size N, such that if X of the samples are successfully tested, then a (1-α) 100% lower confidence bound (LCB) for the probability of success (p) is at least γ. Prior to the test, it must be clearly understood what constitutes a passing trial. The values of 1-α and γ should also be agreed upon before testing commences. If the system is one in which the probability of success is desired to be high, then the value of γ should also be close to one. The value of 1-α, reflecting the confidence level, should also be fairly high. In studies where the cost of conducting each trial is expensive, it may be critical to keep the number of tests to a minimum. One way to minimize testing is to incorporate a zero-failure policy whereby each trial must be successful in order for the reliability standard to be met.

Now consider a specific problem of this nature relating to a zero-failure study of the reliability of armor packages in defeating a prescribed threat. For this study, both 1-α and γ are set at 90%, meaning that based upon the results of N successful trials, we wish to conclude with 90% confidence that the armor package is capable of defeating the threat with a minimal probability of 90%. The question of interest is "What is the value of N that, if no failures are observed, allows us to meet this 90/90 reliability specification?"

The answer to this question is determined by examining how the true but unknown success probability is estimated. This straightforward estimation problem has been the subject of research for nearly two centuries. Many solutions have been proposed to this problem which continues to draw interest today.

2. Interval Estimation of Binomial Proportions

The armor reliability problem is tantamount to estimating the parameter of a binomial distribution. The binomial distribution is used to model the number of successes (X) out of N Bernoulli trials when the probability of success is p. If all N trials are successful, then $X = N$ and the maximum likelihood estimate for p is $\hat{p} = 1$. This point estimate is not very enlightening, since it conveys no information on the variability associated with $\hat{p}$. As long as no failures are observed, the same value for $\hat{p}$ is returned whether 2 or 2,000,000 trials are conducted. Intuitively, as the value of N increases, we are much more likely to believe that the

true value of *p* is close to 1. What we prefer to report with some degree of confidence is a range of plausible values for *p*, i.e., an interval estimate for the true probability of success.

When *p* is believed to be high, there is usually little interest in placing an upper confidence limit on its value. However, a lower limit, known as an LCB for *p*, is useful since the LCB represents a conservative estimate on the probability of success. Because the LCB is based on the random sample of Bernoulli trials, it too is a random variable. The LCB is a function of X and *N* which satisfies the probability statement

$$1-\alpha = P\left(LCB\left(X;N\right) \le p\right). \tag{1}$$

Many authors have proposed methods for calculating LCBs (and confidence intervals) for the binomial parameter *p*. In the ensuing subsections, we introduce three of them, and calculate the required sample size for a zero-failure test that satisfies the 90/90 reliability specification.

2.1 The Clopper-Pearson Method

The Clopper-Pearson (1934) method for binomial confidence intervals is popular for its relative ease to calculate. In general, the confidence interval limits, $\left(L_{CP}, U_{CP}\right)$, are solutions to the statements $\sum_{i=X}^{N}\binom{N}{i}L_{CP}^{i}\left(1-L_{CP}\right)^{N-i}=\frac{\alpha}{2}$ and $\sum_{i=0}^{X}\binom{N}{i}U_{CP}^{i}\left(1-U_{CP}\right)^{N-i}=\frac{\alpha}{2}$. If at least one success and one failure are observed among the samples, both endpoints of the interval can be expressed as functions of percentiles from *F* distributions. For an LCB, with $X = N$, the calculation simply reduces to $L_{CP} = \sqrt[N]{\alpha}$. Clopper-Pearson intervals are often referred to as "exact" intervals since they are derived from exact probability statements and not any distributional approximations. As such, the Clopper-Pearson is often touted in introductory statistics textbooks.

To satisfy the 90/90 reliability specification using the Clopper-Pearson method, we seek the minimum value of *N* which satisfies $.90 \le .10^{1/N}$. Taking the logarithm of both sides of this inequality, we have $\ln\left(.90\right) \le \frac{\ln\left(.10\right)}{N}$ which leads to the solution $N \ge \ln\left(.1\right)/\ln\left(.9\right) = 21.85$. Since *N* must be an integer, the number of zero-failure trials required to meet the 90/90 specification under the Clopper-Pearson method is rounded up to 22.

2.2 Wilson Score Method

The method developed by Wilson (1927) is based on an inversion of the score test for *p*, and results in a more complex formula for the limits of the confidence interval:

$$\left(\frac{X}{N} + \frac{z_{\alpha/2}^2}{2N} \pm z_{\alpha/2} \sqrt{\frac{\frac{X}{N}\left(\frac{N-X}{N}\right) + \frac{z_{\alpha/2}^2}{4N}}{N}} \right) \Bigg/ \left(1 + \frac{z_{\alpha/2}^2}{N} \right), \tag{2}$$

where z_δ is the value having an area of δ to its right under the standard normal curve. However, with no failures, the formula for an LCB reduces to $LCB = \frac{N}{N + z_\alpha^2}$.

To satisfy the 90/90 reliability specification using the Wilson score method, we seek the minimum value of *N* which satisfies $.90 \le \frac{N}{N + z_{.10}^2}$. The value of $z_{.10} \approx 1.2816$ can be found in most statistics textbooks. After a few algebraic manipulations, we get a solution of $N \ge 9z_{.1}^2$, or $N \ge 14.78$. Since *N* must be an integer, the number of zero-failure trials required to meet the 90/90 specification under the Wilson score method is rounded up to 15.

2.3 Jeffreys Method

The final interval construction method that we examine is Jeffreys method, first proposed by Rubin and Schenker (1987). This method is based on a Bayesian estimate for the success probability, whereby *p* is not considered a fixed unknown parameter but rather a random variable. The prior distribution proposed by Jeffreys method is a beta distribution with both parameters set to 0.5; and the posterior distribution of *p* is given by a beta distribution with parameters $X + 0.5$ and $N - X + 0.5$. Note that these two parameters are the number of successes and the number of failures, both of which are then increased by 1/2. The confidence interval endpoints are the values within the support of the posterior distribution that define the lower and upper $\alpha/2$ percentiles. For the construction of an LCB, we have

$$LCB = BetaCDF^{-1}\left(\alpha, X + \tfrac{1}{2}, N - X + \tfrac{1}{2}\right). \tag{3}$$

In our specific case of $X=N$, the LCB is the value within the support of a beta distribution with parameters $N + 0.5$ and 0.5 whose cumulative distribution function equals α.

To satisfy the 90/90 reliability specification using Jeffreys method, we seek the minimum value of *N* which satisfies the inequality

$$.90 \le BetaCDF^{-1}\left(.10, N + \tfrac{1}{2}, \tfrac{1}{2}\right). \tag{4}$$

Although a closed-form solution is not tenable, analytic software capable of calculating the inverse of a beta cumulative distribution function (CDF) (e.g., MATLAB) is used to obtain $N \ge 12.58$. Since *N* must be an integer, the number of zero-failure trials required to meet the 90/90 specification under Jeffreys method is 13.

3. Comparison of Different LCB Methods

A (1-α) 100% confidence interval is usually interpreted as a range of values which contains the true but unknown parameter value with probability 1-α. So, for example, if 1000 independent data sets are used to generate 95% confidence intervals for some parameter, we can expect about 950 of them to contain that parameter value p. However, when a confidence interval is based on approximate distributional theory and/or discrete distributions, the actual "coverage probability," or proportion of time that the interval contains p, might not equal 1-α. Furthermore, the coverage probability may depend upon the value of the parameter p. Coverage probabilities exceed the nominal coverage probability 1-α when the confidence intervals are unnecessarily wide; we say that such intervals are conservative. On the other hand, if the confidence intervals tend to be too narrow, the coverage probabilities will be less than the nominal value (1-α).

To compare various confidence intervals, we need to examine their coverage probabilities. Several authors (Ghosh, 1979; Blyth and Still, 1983; Agresti and Coull, 1998; Brown et al., 2001) have done just this in studying the coverage probabilities of two-sided confidence intervals for the binomial parameter p. Agresti and Coull, in particular, found that the exact method is highly conservative and noted that the Wilson score method results in actual coverage probabilities near the nominal level; however, they did not included Jeffreys method in their paper. Cai (2005) points out that good performance in terms of two-sided interval coverage probabilities does not necessarily guarantee that a method will perform similarly for one-sided intervals. Cai compared the Jeffreys and Wilson score methods, along with other candidate methods, for the coverage probabilities of 99% upper confidence bounds.

In this section, we focus on the coverage probabilities of 90% lower confidence bounds for the binomial parameter using the three methods outlined in section 2. Monte-Carlo simulation was utilized to estimate the coverage probabilities. For a given N and p, a large number of binomial observations are randomly generated. Then using each of the three methods, the LCBs are calculated. The estimated coverage probabilities are equal to the frequency with which the LCBs are less than p.

Figure 1 shows the actual coverage probability (based on 100,000 simulated binomial draws) as a function of the probability of success, p, for the Clopper-Pearson, Wilson score, and Jeffreys intervals when the sample size is 10, 25, and 50. The probability of success is limited to values above 0.7 since the application of this study is to a system whose p is relatively high.

The most obvious feature of each of these plots is the saw-toothed relationship between p and coverage probability, an artifact of the discreteness of the binomial distribution. Also worth noting in figure 1 is that for any choice of α, N, and LCB construction method, there is an entire interval of values $\left[p^*, 1 \right]$ for which the coverage probability is 100%. It so happens that p^* is

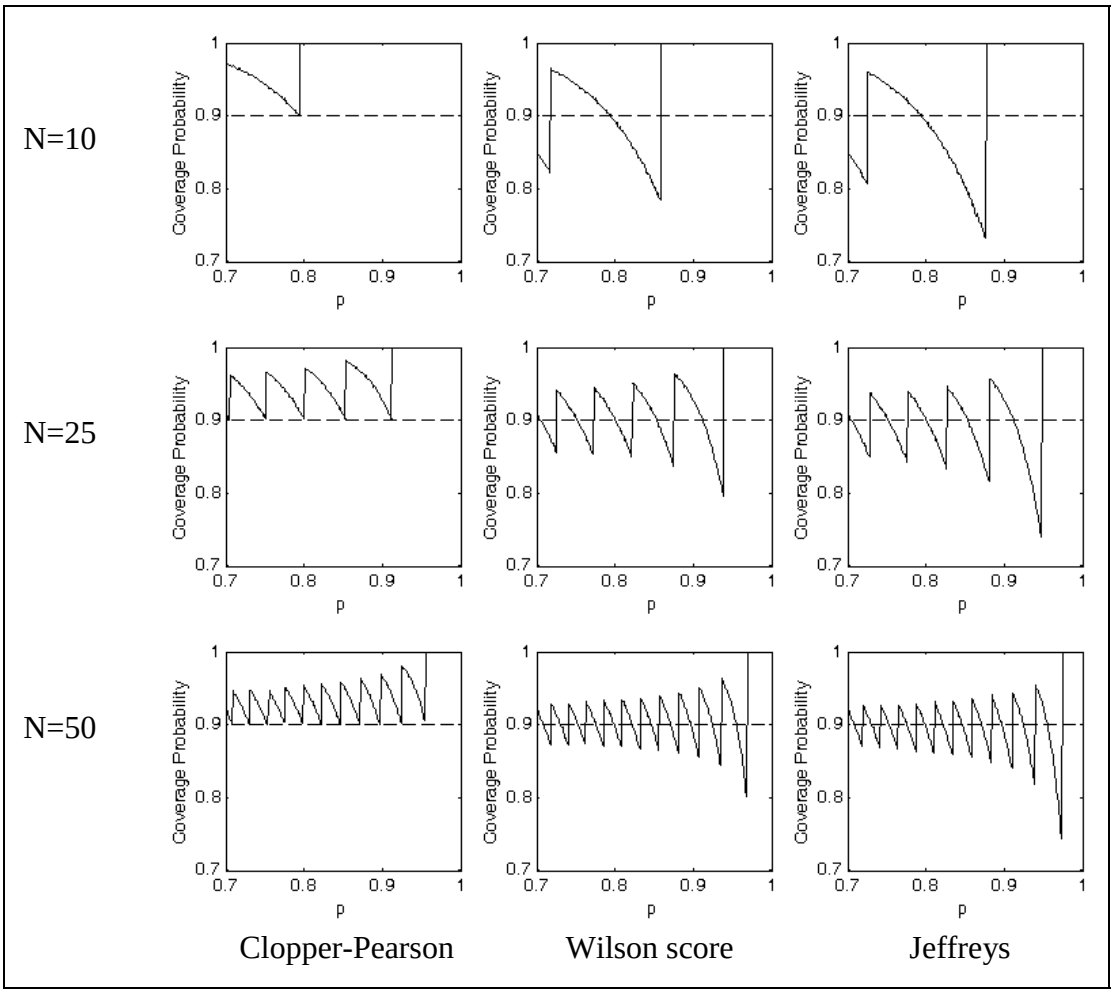

Figure 1. A comparison of coverage probabilities vs. true binomial parameter for the nominal 90% LCB.

the LCB for a zero-failure experiment; e.g., for the Clopper-Pearson method, $p^*_{CP} = \sqrt[N]{\alpha}$; for Wilson score, $p^*_{WS} = N/\left(N + z^2_\alpha\right)$; and for Jeffreys, $p^*_J = BetaCDF^{-1}\left(\alpha, N + .5, .5\right)$. For most practical values of α and N, including those in figure 1, p^*_{CP} is smaller than either p^*_{WS} or p^*_J.

With coverage probabilities always exceeding (or at) the nominal value, the Clopper-Pearson method is clearly the most conservative of the three methods. Both the Wilson score and Jeffreys LCBs have coverage probabilities which can be greater than or less than the nominal coverage probability depending upon the true value of *p*. The oscillation about 90% nominal coverage is slightly less under the Wilson score method.

Since the selection of a "best" method for constructing LCBs is dependent on *p*, the concept of mean coverage probability over the range of possible values of *p* allows us to judge which method on average is preferred. Sampling values of *p* from a uniform distribution over the range

of interest (i.e., [0.7, 1.0]) and then averaging the sample of associated coverage probabilities is the Monte-Carlo equivalent of integrating the *p*-versus-coverage-probability curve. This is exactly what is done to produce figure 2a for $N = 5{:}5{:}100$ and generating 250,000 values from each distribution. We see that on average all three methods are conservative to some degree. The most conservative method is Clopper-Pearson, followed by Wilson score and then Jeffreys.

Instead of assuming that the values of *p* are equally likely to be between 0.7 and 1.0, one can assume that *p* follows some other distribution. A natural choice is the beta distribution since its support is on the interval [0, 1]. In figure 2b, we assume that the success probability is a beta random variable with parameters 13.6 and 2.4. These parameters were chosen to match the first two moments of a uniform (0.7, 1.0) distribution. This results in slightly more conservative coverage of Wilson score and Jeffrey LCBs for very small sample sizes (N = 5); however, as *N* increases, the bias in coverage probability for these two LCBs is more quickly reduced.

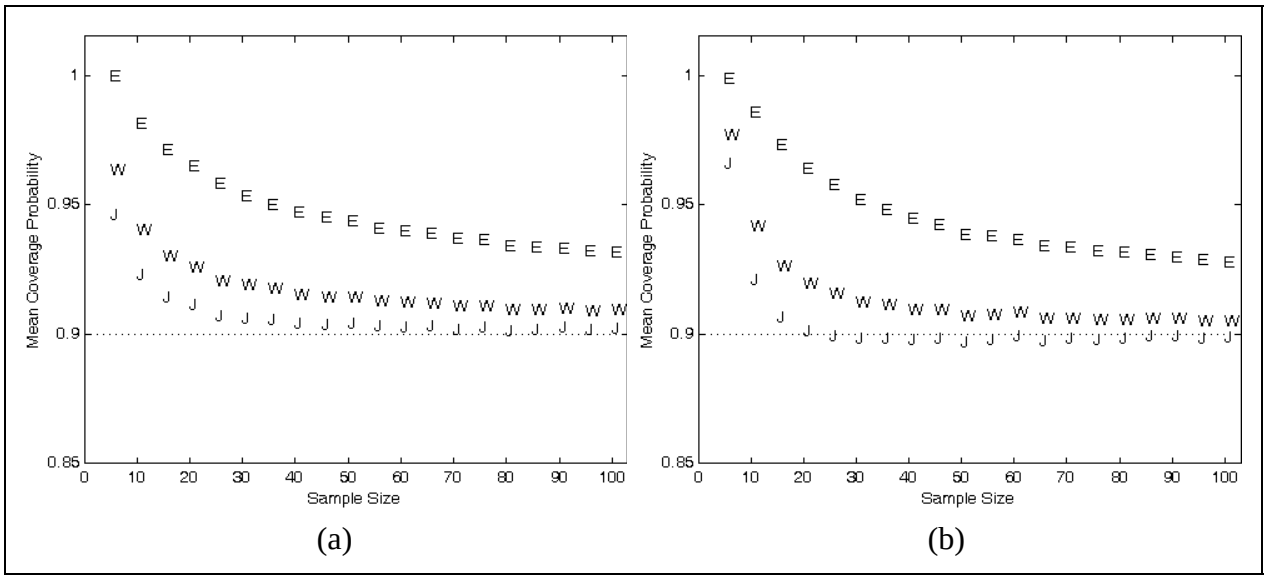

Figure 2. Mean coverage probability as a function of *N:* for the nominal 90% Clopper-Pearson (E), Wilson score (W), and Jeffreys (J) LCBs, when *p* has (a) a uniform distribution on (.7, 1) and (b) a beta distribution with parameters 13.6 and 2.4.

These two figures corroborate the ultra-conservative nature of Clopper-Pearson LCBs and lead us to conclude that a 22-trial study is unnecessary. As long as no failures are observed, both a 15-trial study using Wilson score LCBs and a 13-trial study using Jeffreys LCBs will allow us to conclude that the 90/90 reliability specification has been met. Because the Wilson score method requires two additional successful trials to reach this same conclusion, it is slightly more conservative.

4. Sample Size Requirements for Nonzero Failure Testing

If the cost of testing is not too excessive, a small number of failures may be permissible among the test trials. In this section, we determine the sample size requirements for tests that permit either 1 or 2 failures, and still allow the 90/90 reliability criteria to be met. We do so for both the Wilson score and Jeffreys methods.

4.1 Wilson Score Method With One Failure

To calculate the necessary sample size, we consider the lower limit of equation 2, setting $X = N - 1$ and $\alpha = 0.10$ to get the inequality

$$.9 \le \frac{\frac{N-1}{N} + \frac{z_{0.9}^2}{2N} - z_{0.9}\sqrt{\frac{\frac{N-1}{N^2} + \frac{z_{0.9}^2}{4N}}{N}}}{1 + \frac{z_{0.9}^2}{N}}. \tag{5}$$

A closed-form solution for N is not tractable, however using trial and error, one can show that $N = 32$ is the minimum sample size which meets the 90/90 reliability specification in a one-failure test using the Wilson score method.

4.2 Wilson Score Method With Two Failures

Similarly, setting $X = N - 2$ in the lower limit of equation 2, we get

$$.9 \le \frac{\frac{N-2}{N} + \frac{z_{0.9}^2}{2N} - z_{0.9}\sqrt{\frac{\frac{2(N-2)}{N^2} + \frac{z_{0.9}^2}{4N}}{N}}}{1 + \frac{z_{0.9}^2}{N}}. \tag{6}$$

To meet a 90/90 reliability specification, the minimum sample size necessary for a two-failure test using the Wilson score method is $N = 47$.

4.3 Jeffreys Method With One Failure

To calculate the necessary sample size, we consider the lower limit of equation 3, setting $X = N - 1$ and $\alpha = 0.10$ to get the equation

$$.9 \le BetaCDF^{-1}\left(.10, N - 1/2, 3/2\right). \tag{7}$$

Again, a closed form expression for the minimum value of N does not exist. However, using statistical software capable of calculating the inverse CDF for a beta distribution, we determine that the minimum sample size necessary to meet a 90/90 reliability specification in a one-failure test using the Jeffreys method is $N = 30$.

4.4 Jeffreys Method With Two Failures

Similarly, setting $X = N - 2$ in the lower limit of equation 3, we get

$$.9 \le BetaCDF^{-1}\left(.10, N - 3/2, 5/2\right). \tag{8}$$

To meet a 90/90 reliability specification, the minimum sample size necessary for a two-failure test using the Jeffreys method is $N = 45$.

To summarize this section, the required sample sizes for 90/90 tests using various failure allowances and LCB construction methods are shown in table 1.

Table 1. Minimum number of trials required to test for a 90/90 reliability standard.

		LCB Method		
		Clopper-Pearson	**Wilson score**	**Jeffreys**
Number of Failures Allowed	0	22	15	13
	1	38	32	30
	2	52	47	45

5. Multistage Sampling Plans

For any of the aforementioned tests, it should be obvious that a 90/90 test may be terminated once the number of allowable failures is exceeded. This may occur if the quality of the product is poor; or if the product is of satisfactory quality but suffers from poor luck during the test. Now consider the stopping of tests prematurely for exceptional quality by utilizing a "multistage" sampling strategy. For example, suppose that one is using the Wilson method with one allowable failure. The test would normally call for a sample size of 32. However, if we observe successes in each of the first 15 trials, then we can stop the test prematurely since at this point the 90/90 reliability specification is met. Such a testing strategy we refer to as a two-stage test. In a three-stage test using the Wilson score method, we stop if

1. Zero failures are observed in the first stage (the first 15 trials);
2. Only one failure is observed among first and second stages (the first 32 trials);
3. Three failures are observed at any point in the test; or

4. All three stages are completed, i.e., all 47 trials have been run.

Graphically, we can display the sampling protocol for a multistage test using the technique of figure 3, which is drawn for our 90/90 test using the Wilson score method. (Note that a one-stage test is synonymous with a zero-failure test.) In this type of chart, the progression of the test is mapped as a function of the number of trials and the number of observed trials. Landing on either a square or circle terminates the test—a square indicates that the number of allowable failures has been exceeded, whereas a circle indicates that the 90/90 reliability specification is met. Landing on a dot means that testing should continue. Sampling protocol charts under the Jeffreys method are not included in this document.

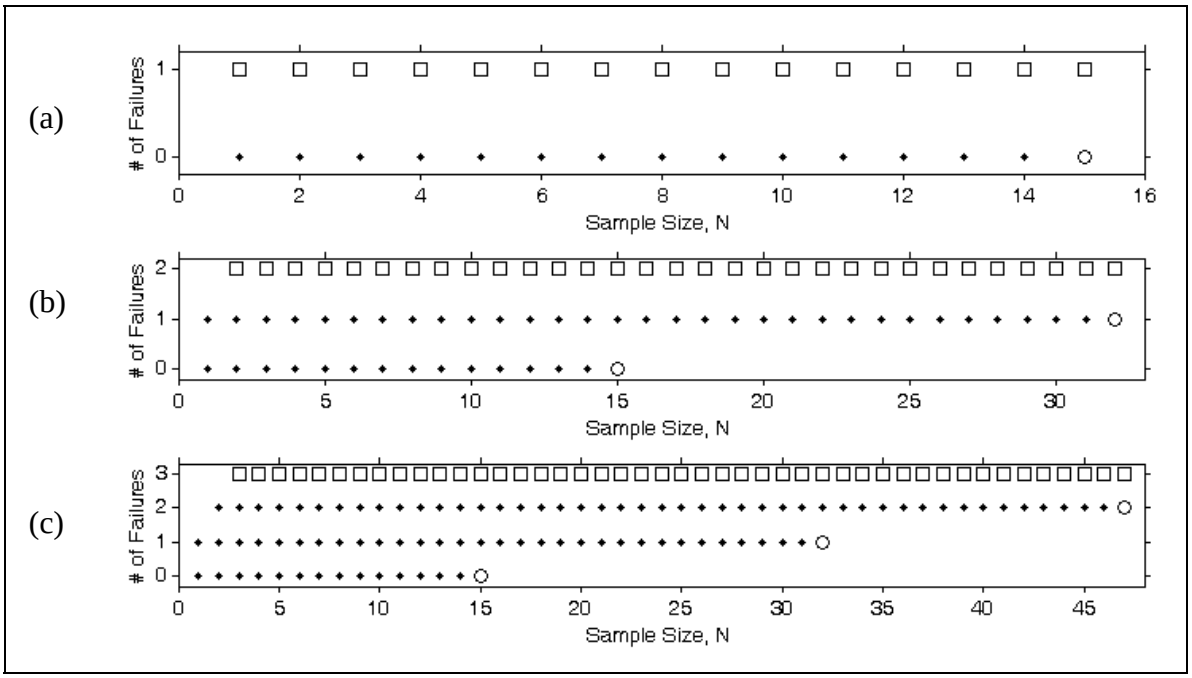

Figure 3. Sampling protocol charts for a 90/90 (a) one-stage test, (b) two-stage test, and (c) three-stage test, each using the Wilson score method for estimating the LCB. Travel starts at (0,0) and moves one unit to the right with each trial, and an additional unit upward if that trial is a failure. If the current test "position" is on a dot, then testing continues. If the current test position is on a square, then testing stops without meeting the 90/90 reliability specification. If the current test position is on a circle, then testing stops with the 90/90 reliability specification satisfied.

5.1 Operating Characteristic Curves

Assuming that the probability of success is known, then the probability of meeting the 90/90 reliability specification can be calculated. For example, in a one-stage test consisting of up to N_1 trials, this is simply the probability that all trials are successes, p^{N_1}. In a two-stage test of up to N_2 trials, the reliability specification is satisfied if either (1) all of the N_1 first-stage trials are successes, or (2) if exactly one of the first-stage trials is a failure AND all of the $N_2 - N_1$ second-stage trials are successes. The probability of this event is

$$
\begin{aligned}
P\left(meet\ spec\right) &= P\left(all\ S\ in\ trials\ 1\ to\ N_1\right) \\
&\quad + P\left(1\ F\ in\ trials\ 1\ to\ N_1\right) \times P\left(all\ S\ in\ trials\ N_1+1\ to\ N_2\right) \\
&= p^{N_1} + \binom{N_1}{1} p^{N_1-1}\left(1-p\right) p^{N_2-N_1+1}
\end{aligned}
\tag{9}
$$

The general formula for the probability of meeting the 90/90 reliability specification in a three-stage test is more complex and not shown here. However, this probability as a function of the percent defective (probability of failure for any individual trial) is plotted in figure 4. As expected, the relationship is decreasing and the probability of successfully meeting the 90/90 reliability criteria increases as more stages (and hence more trials) are added. These plots are akin to operating characteristic curves frequently shown in quality-control circles, and are helpful in providing an *a priori* estimate for the probability of a successful test.

For example, figure 4a shows that even if the true success probability is 95% (5% defective rate of 5%), there is only a 46% chance that a 15-trial one-stage test using Wilson score LCBs will result in all successes and meet the 90/90 reliability specification. The same material in a 13-trial one-stage using Jeffrey LCBs only has a 51% chance of meeting the specification. This highlights the risk of conducting a small-sample test to pass material at a high level of performance. Use of a two- or three-stage test will improve the chance that high quality material is found to meet the specification, but at the price of nearly doubling (or tripling) the number of trials.

5.2 Sample Size Expectation

Another important characteristic of a sampling protocol is the expected number of samples until test termination. Because a multistage test can be stopped early, the number of trials conducted until the test is terminated is a random variable. Therefore, we can calculate its expected value.

For example, denoting the number of trials conducted in a one-stage test by M_1, its expectation is $E\left(M_1\right) = \sum_{i=1}^{15} i \bullet P\left(M_1 = i\right)$. Note that if $1 \le M_1 \le N_1 - 1$, then the test is stopped early because trial M_1 is a failure while all prior trials were successes. If $M_1 = N_1$, then each of the first $N_1 - 1$ trials was a success. So,

$$
\begin{aligned}
E\left(M_1\right) &= \sum_{i=1}^{N_1-1} i \bullet P\left(M_1 = i\right) + N_1 \bullet P\left(M_1 = N_1\right) \\
&= \sum_{i=1}^{N_1-1} i \bullet P\left(all\ S\ in\ trials\ 1\ to\ i-1,\ then\ F\right) + N_1 \bullet P\left(all\ S\ in\ trials\ 1\ to\ N_1 - 1\right)
\end{aligned}
\tag{10}
$$

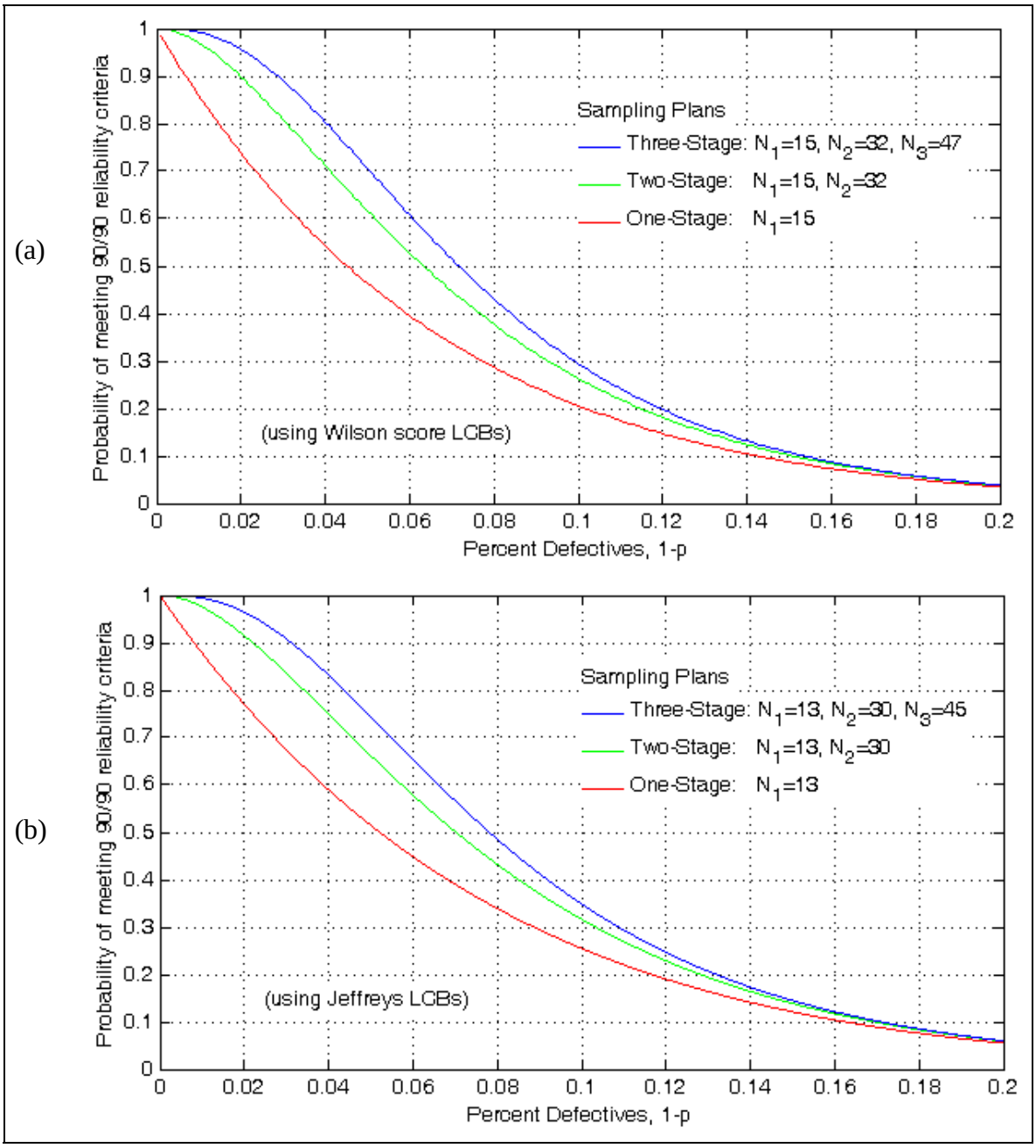

Figure 4. Relationship between percentage of defective items in the population and the probability of meeting the 90/90 reliability specification using (a) Wilson score and (b) Jeffreys methods of computing the LCB for reliability.

Each of the probabilities in equation 9 can be evaluated using the binomial distribution, leading to

$$E(M_1) = (1-p)\sum_{i=1}^{N_1-1} ip^{i-1} + N_1 p^{N_1-1}. \qquad (11)$$

Substituting for the summation of a finite series in the left addend, we get

$$E\left(M_1\right)=\left(1-p\right)\left(\frac{1-p^{N_1}}{\left(1-p\right)^2}-\frac{N_1p^{N_1-1}}{1-p}\right)+N_1p^{N_1-1}=\frac{1-p^{N_1}}{1-p}. \tag{12}$$

While the details are omitted here for brevity, it can be shown that the expected number of trials conducted in a two-stage test, M_2, is

$$E\left(M_2\right)=\frac{2\left(1-p^{N_1}\right)}{1-p}-N_1p^{N_2-1}; \tag{13}$$

and the expected number of trials for a three-stage test equals

$$E\left(M_3\right)=3\left(\frac{1-p^{N_1}}{1-p}\right)-2N_1p^{N_2-1}-\frac{N_1}{2}\left(N_3\left(N_1-1\right)\left(1-p\right)+\left(2N_2-N_1-1\right)p\right)\left(1-p\right)p^{N_3-3}. \tag{14}$$

Figure 5 displays the relationship between percent defectives and the expected number of samples for various staged sampling protocols and LCB methods. In one-stage tests, as the percent defectives decreases we are more likely to carry out the full test of N_1 trials. For two- and three-stage tests, this same principle applies up to a point at which the decreasing percent defectives results in the increased application of early stopping rules.

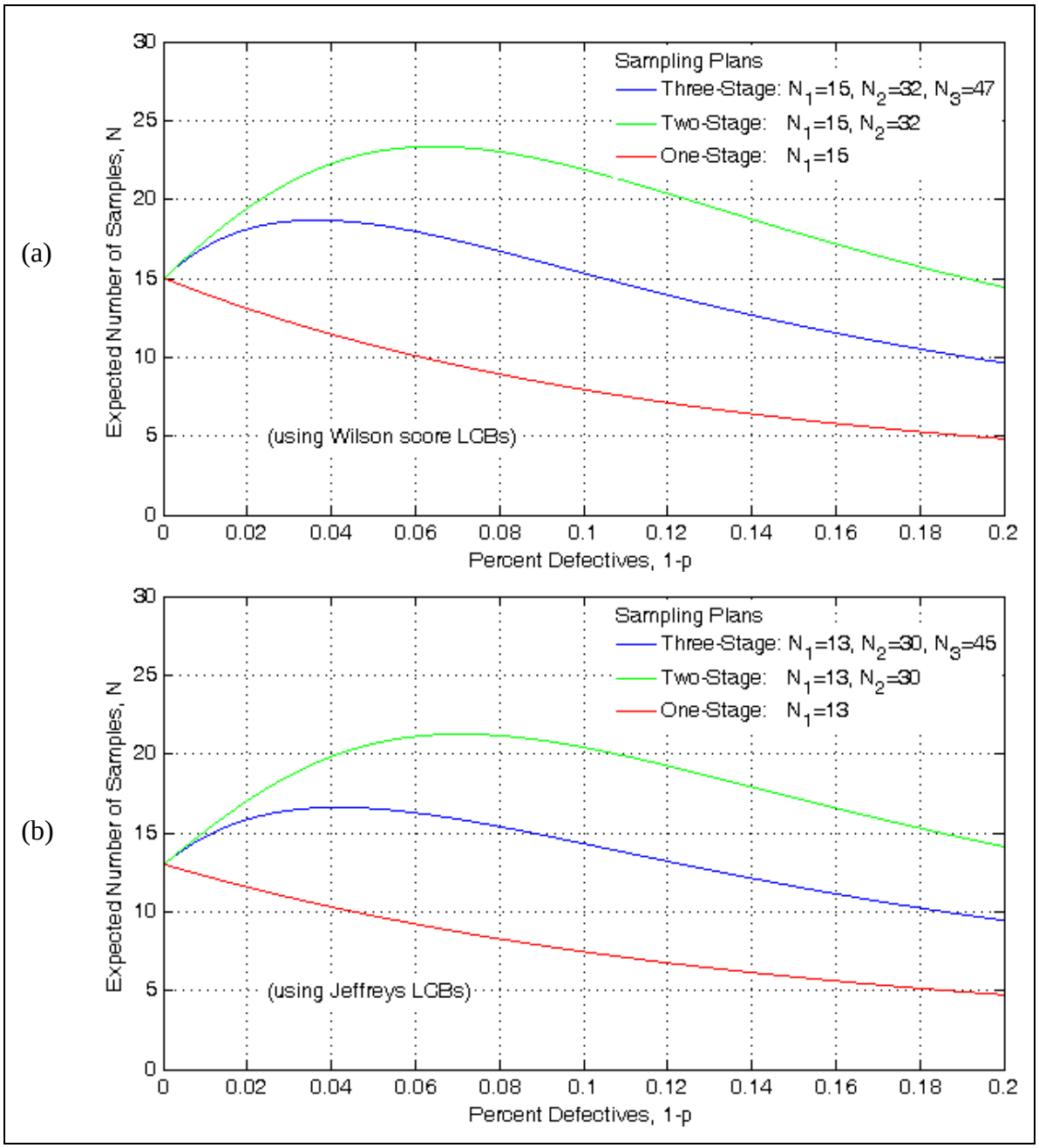

Figure 5. Percentage of defective items in the population vs. expected number of samples under several sampling protocols using (a) the Wilson score and (b) Jeffreys methods of constructing LCBs.

6. Summary

Over the years, and even to this day, many statistics texts have advocated use of the Clopper-Pearson method for interval estimation because of its "exactness." Hence, many zero-failure acceptance tests have been conducted with 22 trials with the intent of showing at least 90% reliability with 90% confidence. This report has shown that this sample size is unnecessarily high. It is possible using the Wilson score method of LCB construction to show this same level of performance with only 15 trials, a savings of 32% in test resources. The savings under the Jeffreys method is even greater—only 13 trials and a 41% cost reduction.

Both the Wilson score and Jeffreys methods do come with some risks. On average, both methods are slightly conservative in terms of their coverage probability; however, there are some values of p for which the actual coverage probability of a 90% Wilson score LCB can be closer to 80%, meaning that the LCBs tend to be too large. Figure 4 reveals that for large-sample tests using a Jeffreys LCB, the coverage probability may be as small as 75% for a limited range of success probabilities near 97%, meaning that there is nearly a 1-in-4 chance that certain high-grade material may not meet the 90/90 specification.

If a conservative estimate of at least 90% reliability is desired from our test, then the actual probability of success for the system should be substantially greater than 90%. Figure 4 shows even when $p = 0.95$, there is only about a 45% chance of confirming that the system is functioning at the desired level of performance when using a one-stage, Wilson score-based test. When $p = 0.99$, this chance increases to about 85%. Because fewer tests are required under a Jeffreys-based test procedure, the probability of meeting the 90/90 specification are higher by about 7% when $p = 0.95$ and 3% when $p = 0.99$. These observations are made to point out that there is a significant risk of a failing to pass even high-quality material under a small-sample test protocol in which no failures are allowed. If we are willing to commit to a larger two-stage test, we stand a much greater chance of showing that good-quality material (e.g., armor packages) meets the reliability specification.

7. References

Agresti, A.; Coull, B. A. Approximate Is Better Than Exact for Interval Estimation of Binomial Proportions. *The American Statistician* **1998**, *52*, 119–126.

Blyth, C. R.; Still, H. A. Binomial Confidence Intervals. *J. of the American Statistical Assoc.* **1983**, *78*, 108–116.

Brown, L. D.; Cai, T. T.; DasGupta, A. Interval Estimation for a Binomial Proportion. *Statistical Science* **2001**, *16* (2), 101–133.

Cai, T. T. One-Sided Confidence Intervals in Discrete Distributions. *J. of Statistical Planning and Inference* **2005**, *131*, 63–88.

Clopper, C.; Pearson, E. S. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika* **1934**, *26*, 404–413.

Ghosh, B. K. A Comparison of Some Approximate Confidence Intervals for the Binomial Parameter. *J. of the American Statistical Assoc.* **1979**, *74*, 894–900.

Rubin, D. B.; Schenker, N. Logit-Based Interval Estimation for Binomial Data Using the Jeffreys Prior. *Sociological Methodology* **1987**, *17*, 131–144.

Wilson, E. B. Probable Inference, the Law of Succession, and Statistical Inference. *J. of the American Statistical Assoc.* **1927**, *22*, 209–212.

NO. OF COPIES	ORGANIZATION
1 (PDF only)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA 8725 JOHN J KINGMAN RD STE 0944 FORT BELVOIR VA 22060-6218
1	DIRECTOR US ARMY RESEARCH LAB IMNE ALC HRR 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIM L 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL CIM P 2800 POWDER MILL RD ADELPHI MD 20783-1197
1	DIRECTOR US ARMY RESEARCH LAB RDRL D 2800 POWDER MILL RD ADELPHI MD 20783-1197
	ABERDEEN PROVING GROUND
1	DIR USARL RDRL CIM G (BLDG 4600)

NO. OF
COPIES ORGANIZATION

ABERDEEN PROVING GROUND

3 USAEC
N DUNN
C TURNER
Z ZIMMER
4120 SUSQUEHANNA AVE
APG MD 21005

1 DIR USAMSAA
RDAM LF
J NIERWINSKI
392 HOPKINS RD
APG MD 21005-5701

33 DIR USARL
(32 HC RDRL CII C
1 CD) B BODT
RDRL SLB D
J COLLINS
L MOSS
RDRL SLB E
P HORTON
RDRL WMM
J BEATTY
RDRL WMM A
M MAHER
RDRL WMM D
R CARTER
E CHIN
RDRL WMM E
T JESSEN
RDRL WML A
B FLANDERS
B OBERLE (CD only)
A THOMPSON
D WEBB (12 CPS)
RDRL WMP
P BAKER
S SCHOENFELD
RDRL WMP D
M KEELE
D KLEPONIS
D PETTY
J RUNYEON
RDRL WMP E
M BURKINS
D HACKBARTH
E HORWATH
K KRAUTHAUSER

INTENTIONALLY LEFT BLANK.